Uso de Ensembles en Problemas con Cambios de Distribución Caracterizables*

Pablo Pérez-Gallego, José Ramón Quevedo, Juan José del Coz
{pablopg,quevedo,juanjo}@aic.uniovi.es

Centro de Inteligencia Artificial, Universidad de Oviedo en Gijón (Spain)

Resumen Entre los métodos de aprendizaje supervisado, los *ensembles* suelen tener un destacado rendimiento en la práctica, sobre todo cuando existe diversidad entre los modelos que combinan. Esa diversidad puede lograrse generando muestras de entrenamiento con ciertas diferencias para cada uno de los modelos. Es decir, los modelos se entrenan con un distribución de los datos distinta de la original. Por ese motivo, la hipótesis de este artículo se basa en que los *ensembles* pueden ser especialmente apropiados en aquellos problemas en los que existen cambios en la distribución y que además dichos cambios se pueden caracterizar. La idea es entrenar los distintos modelos con distintos conjuntos de entrenamientos generados basándose en los cambios esperados en la distribución. Como caso de estudio, nos centraremos en los problemas de cuantificación binaria y presentaremos métodos *ensembles* de dos algoritmos de cuantificación. Los resultados experimentales muestran que los *ensembles* superan a los algoritmos individuales, incluso cuando se usan estrategias de combinación simples.

Palabras clave: Cambios en la distribución, *Ensembles*, Cuantificación

1. Introducción

Los *ensembles* se basan en construir un meta-modelo que resulta de combinar, usando una determinada regla, un conjunto de modelos, de forma que la diversidad existente en los mismos permite obtener una solución consensuada y, en general, con más efectividad que un modelo único, aunque esto no siempre está garantizado [7]. Intuitivamente, se trata de una técnica muy presente en el día a día para el ser humano; un conjunto de opiniones es más enriquecedora que una opinión por separado, en especial si existe diversidad entre las mismas.

La asunción principal del aprendizaje supervisado es que la distribución desconocida $P(x, y)$ de la que se extraen los ejemplos no cambia entre entrenamiento y test. No obstante, esta asunción de partida se ve a menudo violada en las aplicaciones reales [13,14]. Este fenómeno se conoce como *dataset shift* y ocurre cuando $P(x, y)$ cambia en el test respecto del entrenamiento [12]. Aunque en ciertos casos caracterizar estos cambios es complicado y dependería de la aplicación abordada, en otros es bastante sencillo, incluso trivial.

* Trabajo financiado parcialmente por el MINECO, proyecto TIN2011-23558.

La intención de este artículo es presentar un nuevo escenario donde la aplicación de *ensembles* resulta adecuada y efectiva. Se trata de aquellos problemas en los que se puede caracterizar cómo son los cambios en la distribución. El objetivo es aprovechar ese conocimiento y utilizarlo durante el entrenamiento. En el contexto del aprendizaje de *ensembles* esto nos puede ayudar, además, a generar diversidad entre los distintos modelos individuales, característica muy deseable que mejora su efectividad [4]. La idea clave del artículo es generar distintas muestras de forma que cada una representa un cambio concreto en la distribución. Este enfoque es diferente a otros que se han propuesto [10] para tratar ciertos problemas en los que la distribución cambia, principalmente en problemas de *concept drift*. Dichos métodos se basan en borrar, modificar o añadir modelos al *ensemble*, principalmente porque el concepto cambia en el tiempo y a priori no sé puede saber cómo va a ser ese cambio. Nuestro enfoque es diferente porque se conocen las características de esos cambios y se puede utilizar para construir el *ensemble* desde el primer momento, sin necesidad de modificaciones posteriores.

Para probar la validez de nuestra idea, la hemos aplicado al problema de la cuantificación binaria [6]. En dicho problema la probabilidad de las clases puede cambiar y el objetivo es precisamente, dado un conjunto de objetos, estimar la prevalencia o proporción de ejemplos de la clase positiva. Como veremos en los resultados experimentales, la precisión de dichas estimaciones mejora con el uso de *ensembles* con respecto a cuantificadores individuales. No obstante, creemos que el interés del artículo va más allá de este hecho, ya que el enfoque seguido puede ser aplicable en otros problemas con cambios en la distribución, siempre y cuando éstos sean caracterizables de algún modo.

2. Problemas con cambios de distribución

En [12] se categorizan y discuten los problemas que presentan cambios en la distribución, o usando el término del artículo, en los que se produce un *dataset shift*. Los problemas de aprendizaje supervisado vienen definidos por un conjunto de atributos (*covariates*), x, una variable de clase, y, y los datos provienen de una distribución de probabilidad conjunta de ambos elementos, $P(x, y)$. En los problemas con cambios en la distribución es importante saber cómo se generan los datos partiendo de la relación entre x e y. En [5] se identifican dos clases de problemas en este sentido, los problemas $\mathcal{X} \rightarrow \mathcal{Y}$, donde la clase y se determina causalmente por el valor de x y los problemas $\mathcal{Y} \rightarrow \mathcal{X}$ donde los valores de x dependen causalmente de y. Un ejemplo del primer caso es el problema de la detección de *spam*, en el que el contenido del correo determina si es *spam* o no. Ejemplos típicos del segundo caso son los problemas de diagnóstico médico, en el que tener una cierta enfermedad y hace que se den unos síntomas x.

La probabilidad conjunta $P(x, y)$ se puede reescribir en función del tipo de problema como $P(y|x)P(x)$ para los problemas $\mathcal{X} \rightarrow \mathcal{Y}$ y como $P(x|y)P(y)$ para los problemas $\mathcal{Y} \rightarrow \mathcal{X}$. El *dataset shift* ocurre cuando alguno de esos elementos cambia entre el entrenamiento y el test, es decir, cuando $P_{ent}(x, y) \neq P_{tst}(x, y)$. Así podemos distinguir los siguientes tipos de problemas:

- *covariate shift*: problemas $\mathcal{X} \rightarrow \mathcal{Y}$, cambia $P(x)$ pero $P(y|x)$ es constante,
- *prior probability shift*: problemas $\mathcal{Y} \rightarrow \mathcal{X}$, cambia $P(y)$ pero no $P(x|y)$,
- *concept shift* (o *drift*): problemas $\mathcal{X} \rightarrow \mathcal{Y}$, cambia $P(y|x)$ pero $P(x)$ no cambia y en problemas $\mathcal{Y} \rightarrow \mathcal{X}$, cambia $P(x|y)$ pero $P(y)$ es constante.

Los métodos de aprendizaje supervisado asumen, normalmente, que la distribución de probabilidad conjunta permanece inalterada. Sin embargo, hay muchas aplicaciones importantes en las que se producen cambios en mayor o menor medida. El interés de este tipo de problemas desde el punto de vista del aprendizaje de *ensembles* es que alguno de esos cambios se pueden caracterizar fácilmente. Eso es especialmente cierto en el caso del *prior probability shift*, que en la literatura se denomina también como el problema de la cuantificación.

2.1. Cuantificación binaria

Dado un conjunto de entrenamiento donde cada ejemplo está etiquetado con una clase $y_i \in \{+1, -1\}$, el objetivo de la cuantificación binaria consiste en inducir un modelo o cuantificador capaz de ofrecer una estimación, p', de la prevalencia real de la clase positiva, p, para un conjunto de ejemplos no etiquetados. La asunción en la que se basan los métodos de cuantificación es que la probabilidad de las clases $P(y)$ cambia entre entrenamiento y test, pero $P(x|y)$ no varía.

A primera vista parece un problema más sencillo que la clasificación, puesto que no es necesario predecir con certeza cada uno de los ejemplos. Se podría pensar que para cuantificar basta con inducir un clasificador, clasificar con él los ejemplos y contar cuántos hay de cada clase. Sin embargo, se ha demostrado que este método, denominado CC (*Classify and Count*), produce malas estimaciones. El motivo principal es que el clasificador del CC se induce asumiendo que la distribución no cambia, cuando por definición del propio problema lo hace. Es más, el rendimiento de CC suele ser muy pobre, al menos en alguno de los dos extremos, bien cuando p tienda a subir mucho, o a bajar mucho [6]. Esto es debido a que el clasificador tendrá un sesgo, bien tenderá a producir falsos positivos o falsos negativos. Si p baja mucho, y el clasificador tiende a producir falsos positivos $p' \gg p$, y al revés, si p sube mucho, y el clasificador tiende a generar falsos negativos $p' \ll p$. Con lo cual, es imposible que CC funcione bien para todo el posible rango en el que varíe la prevalencia.

En este artículo trabajaremos con dos métodos de cuantificación que siguen enfoques muy diferentes. En primer lugar, el método AC (*Adjusted Count*) propuesto por Forman para reducir el sesgo que el cuantificador hereda del clasificador [6]. Se basa en que, usando el *tpr* (*true positive rate*) y el *fpr* (*false positive rate*) del clasificador, se puede establecer una relación de la prevalencia estimada, p', en función de la real, p:

$$p'(p) = tpr \cdot p + fpr \cdot (1 - p). \tag{1}$$

Despejando podemos escribir la prevalencia real en función de la estimada:

$$p = \frac{p'(p) - fpr}{tpr - fpr}. \tag{2}$$

Luego bastaría con 1) entrenar un clasificador, 2) estimar su *tpr* y *fpr*, 3) clasificar y contar los ejemplos del conjunto de test para obtener p' y 4) obtener la prevalencia real p substituyendo los valores calculados en (2). Es importante remarcar que teniendo en cuenta que se asume que $P(x|y)$ no cambia, las estimaciones del *tpr* y *fpr* son independientes de los cambios en la distribución posteriores. Es decir, en teoría AC debería producir cuantificaciones perfectas siempre que se mantenga la asunción de que $P(x|y)$ no cambia. Sin embargo, en la práctica eso no ocurre, porque dicha asunción no se cumple y/o porque las estimaciones del *tpr* y *fpr* no son perfectas. En realidad, la corrección de AC es aplicable a cualquier cuantificador que se base en las clasificaciones de un modelo, como en los casos de los cuantificadores basados en árboles de decisión [11], en el vecino más próximo [2], en clasificadores estructurados que optimizan medidas de cuantificación [1] o incluso en clasificadores probabilísticos, aunque en este caso la corrección está adaptada al dominio probabilístico [3].

El segundo método usado, HDy [9], es totalmente diferente a los métodos comentados ya que no se basa en clasificar, contar y ajustar. Su idea radica en buscar similitud entre distribuciones. En concreto, trata de buscar la prevalencia p que haría más similar la muestra de entrenamiento, modificada en función de p y respetando que $P(x|y)$ no cambia, respecto a la distribución del conjunto de test que tenemos que cuantificar. HDy utiliza las predicciones de un modelo probabilístico inducido sobre el conjunto de entrenamiento para representar ambas distribuciones. A través de un parámetro b se particiona el rango [0..1] en b trozos y se construye un histograma donde cada ejemplo de la muestra se lleva a una partición b_i en función de su predicción. La métrica utilizada para medir la similitud entre dos histogramas es la *Distancia de Hellinger*:

$$HD(V,U) = \sqrt{\sum_{i=1}^{b} \left(\sqrt{\frac{|V_{y,i}|}{|V|}} - \sqrt{\frac{|U_{y,i}|}{|U|}} \right)^2}, \tag{3}$$

donde V es el conjunto de entrenamiento, $|V|$ el número de ejemplos que tiene y $|V_{y,i}|$ el número de esos ejemplos cuya predicción pertenece al trozo b_i. U, $|U|$ y $|U_{y,i}|$ se refieren de forma equivalente al conjunto de test.

La estrategia para hacer similar la muestra de entrenamiento respecto a la de test consiste en ir variando su histograma mediante la siguiente fórmula:

$$\frac{|V_{y,i}|}{|V|} = \frac{\left|V^p_{t,y,i}\right|}{|V^p_t|} \cdot p + \frac{\left|V^n_{t,y,i}\right|}{|V^n_t|} \cdot (1-p) \tag{4}$$

donde $|V^p_t|$ es el número de ejemplos de V que pertenecen a la clase positiva y $\left|V^p_{t,y,i}\right|$ el número de ejemplos positivos cuya predicción cae en el trozo b_i. $|V^n_t|$ y $\left|V^n_{t,y,i}\right|$ se refieren análogamente a la clase negativa. Estos componentes se precalculan. La estrategia consiste en ir variando p en el rango [0..1] en pequeños incrementos, obtener el histograma resultante usando (4) y calcular la distancia Hellinger respecto al conjunto de test mediante (3). El valor de p que haga mínima esa distancia será la prevalencia retornada. Nótese que al variar todos los trozos uniformemente usando p se está respetando que $P(x|y)$ no cambia.

3. Ensembles para problemas con cambios de distribución caracterizables

La idea central del artículo se basa en que en aquellos problemas, como *covariate shift* o cuantificación, en los que sabemos cómo cambia la distribución, somos capaces de caracterizarla o al menos se hace una asunción previa de cómo son esos cambios, podemos aprovecharnos de ello para diseñar un *ensemble* con una diversidad apropiada. Como en todo *ensemble* habrá tres fases: 1) generación de las diferentes muestras diversas siguiendo el cambio esperado, 2) entrenamiento de un modelo por cada muestra, y 3) combinación de las estimaciones individuales para generar la predicción final.

El paso clave obviamente es el primero, y debe adaptarse a cada tipo de problema de *dataset shift*. En este artículo vamos a mostrar una adaptación simple para cuantificación binaria. En otros casos, por ejemplo para *covariate shift* en donde lo que cambia es $P(x)$, la adaptación puede requerir de algún conocimiento adicional sobre la aplicación en particular. En el caso de la cuantificación binaria es más sencillo. Recordemos que partimos del supuesto de que los problemas de cuantificación son del tipo $\mathcal{Y} \rightarrow \mathcal{X}$ y de que se producen cambios en $P(y)$ pero que $P(x|y)$ se mantiene constante. El objetivo es generar muestras que, sabiendo el cambio esperado, lo representen a priori y así consigan adaptarse mejor al mismo. Es decir, no se basan en el acierto sobre ejemplos o muestras particulares, aunque podría añadirse esa característica en el futuro, sino solamente en el cambio en la distribución esperado.

El proceso es el siguiente para cada muestra generada. En primer lugar, se selecciona aleatoriamente la prevalencia p que va a tener la muestra. Posteriormente, se realiza un muestreo con reemplazamiento entre los ejemplos de la clase positiva hasta que se obtienen los ejemplos para la prevalencia establecida. Esto garantiza que $P(x|y)$ se mantiene constante. Se repite el mismo proceso para la clase negativa, pero en este caso su prevalencia será $1 - p$. Variando la prevalencia de cada una de las muestras generadas se consigue la diversidad deseada. El procedimiento se repite hasta obtener el número de muestras definido.

El siguiente paso es entrenar el algoritmo de cuantificación base sobre cada muestra generada. En el caso de que el algoritmo emplee la corrección (2) que define AC es necesario estimar el *tpr* y el fpr sobre la muestra. Un detalle importante en el proceso de generación de las muestras es que deben tener un número total de ejemplos adecuado para realizar buenas estimaciones. En nuestro caso, las muestras generadas son del mismo tamaño que el conjunto de datos original.

Finalmente, dado un conjunto de test sin etiquetar, se aplicaría cada uno de los modelos del *ensemble* obteniéndose una estimación de la prevalencia de la clase positiva. Para realizar la estimación final hay que aplicar una función de combinación. En nuestro caso hemos utilizado la media, pero creemos que existe un amplio margen para diseñar métodos de combinación más potentes.

En nuestra opinión la aplicación de *ensembles* al problema de la cuantificación presenta incluso más ventajas que en otro tipo de problemas, como la clasificación. La ventaja que supone emplear una combinación de modelos en lugar de emplear un único modelo, que puede que no sea particularmente bueno,

se da para todos los problemas a los que se aplique esta técnica. No obstante, en el caso de la cuantificación hay ventajas adicionales. En primer lugar, se sabe cómo generar diversidad. Y en segundo lugar, cuando el *ensemble* usa un cuantificador base que emplee la corrección (2), la ventaja de usar varios modelos es aún mayor. La corrección, aunque teóricamente lleva a cuantificar exactamente, es muy peligrosa en la práctica. En el momento en que la estimación del tpr y fpr sea mala, la corrección provocará cambios inapropiados. Como se verá en los resultados experimentales, el método AC puede producir resultados muy buenos en algunos problemas, pero muy malos para otros precisamente por este motivo, porque depende no sólo de la calidad del clasificador, sino sobre todo de la calidad de la corrección. El uso de *ensembles* reduce ambos riesgos. Es decir, que no solamente la diversidad debe producir ventajas, sino también la propia estrategia de combinar varios modelos reduce más riesgos que en un problema de clasificación típico.

4. Resultados experimentales

El objetivo de los experimentos realizados era comprobar si realmente los *ensembles* pueden ser una técnica útil en el contexto de un problema, como es la cuantificación, en el que la distribución de los datos puede llegar a cambiar de una manera brusca. Para ello se compararon el CC (como *baseline*) y dos algoritmos de cuantificación, AC y HDy, con sus correspondientes versiones usando *ensembles*, EAC y EHDy. La elección de AC y HDy se debió a que usan enfoques muy diferentes, lo que aumentaba la amplitud del estudio. Por un lado, AC como algoritmo emblemático y representante de los algoritmos de cuantificación que se basan en la construcción de un clasificador, cuyas predicciones se corrigen posteriormente usando el tpr y el fpr. Por otro lado, HDy, que aunque también usa un clasificador, se basa en representar la distribución de las muestras de ejemplos y no en contar y corregir.

Para realizar los experimentos se emplearon 32 conjuntos de datos, concretamente todos los utilizados en los artículos [9,2], salvo iris.1, acute.a y acute.b (por ser triviales, se obtienen cuantificadores perfectos) y coil, lettersG y lettersH por requerir demasiado tiempo para entrenar los sistemas. Entre los conjuntos elegidos hay varios que son originalmente binarios y otros son versiones binarizadas de un conjunto multiclase. Así por ejemplo, el conjunto iris.2 es un conjunto binario en el que clase 2 original es la clase positiva, formando el resto de ejemplos la clase negativa.

Tanto EAC como EHDy se programaron para que emplearan exactamente los mismos modelos. Lo mismo ocurre con CC, AC y HDy: emplean el mismo clasificador. La representación usada por HDy se generó mediante 8 trozos o *bins*. Con otros valores de ese parámetro HDy obtenía similares resultados (con 4 *bins*) o peores (12, 20 *bins*). Dado que HDy emplea las predicciones de un clasificador para representar las muestras, se optó por usar clasificadores base de salida probabilística para homogeneizar el cálculo de sus histogramas ya que su salida está acotada en [0.,1]. En concreto se utilizaron Naïve Bayes (NB),

Tabla 1. Error cuadrático medio usando regresión logística como clasificador base

conjuntos	CC	AC	EAC	HDy	EHDy
balance.1	0.0031	0.0025	**0.0021**	0.0014	**0.0012**
balance.2	0.1309	0.1833	**0.1328**	0.3708	**0.1530**
balance.3	0.0021	**0.0010**	**0.0010**	**0.0006**	**0.0006**
breast-cancer	0.0008	0.0007	**0.0006**	**0.0004**	**0.0004**
cmc.1	0.0436	0.0123	**0.0114**	0.0118	**0.0105**
cmc.2	0.0429	0.0164	**0.0148**	0.0109	**0.0100**
cmc.3	0.0521	0.0318	**0.0238**	0.0184	**0.0156**
ctg.1	0.0056	0.0016	**0.0013**	**0.0006**	0.0007
ctg.2	0.0092	0.0013	**0.0010**	**0.0010**	0.0011
ctg.3	0.0020	**0.0011**	0.0013	0.0015	**0.0014**
diabetes	0.0211	0.0102	**0.0066**	0.0057	**0.0051**
german	0.0274	0.0091	**0.0085**	**0.0070**	0.0075
haberman	0.0777	0.0780	**0.0662**	0.0966	**0.0735**
ionosphere	0.0152	**0.0068**	0.0108	**0.0131**	0.0185
iris.2	0.0448	0.0677	**0.0457**	0.0329	**0.0189**
iris.3	0.0018	0.0017	**0.0009**	0.0017	**0.0011**
mammographic	0.0168	0.0101	**0.0068**	0.0048	**0.0044**
pageblocks.5	0.0078	**0.0046**	0.0052	0.0016	**0.0010**
phoneme	0.0253	**0.0017**	0.0018	**0.0010**	**0.0010**
semeion.8	0.0070	**0.0017**	0.0046	0.0054	**0.0037**
sonar	0.0230	0.0357	**0.0187**	0.0362	**0.0220**
spambase	0.0079	0.0021	**0.0003**	**0.0002**	**0.0002**
spectf	0.0393	0.0424	**0.0278**	0.0549	**0.0332**
tictactoe	0.0480	0.0365	**0.0289**	0.0203	**0.0172**
transfusion	0.0471	0.0293	**0.0194**	**0.0214**	0.0230
wdbc	0.0066	0.0037	**0.0031**	**0.0023**	0.0024
wine.1	0.0042	0.0035	**0.0026**	0.0024	**0.0021**
wine.2	0.0094	0.0089	**0.0041**	0.0058	**0.0038**
wine.3	0.0016	0.0021	**0.0015**	0.0009	**0.0008**
wine-quality-red	0.0234	0.0082	**0.0069**	0.0051	**0.0047**
wine-quality-white	0.0333	0.0030	**0.0023**	0.0017	**0.0015**
yeast	0.0308	0.0088	**0.0072**	0.0043	**0.0039**
Ranking Medio	4.5625	3.6875	2.5938	2.4688	1.6875

por ser el clasificador probabilístico más típico, regresión logística (RL), para emplear modelos lineales, y SVM con un kernel RBF (gaussiano) con salida probabilística, para obtener modelos no lineales. En el caso de los dos últimos, su parámetro de regularización (C) se estableció mediante una búsqueda en $C \in [10^{-3}, \ldots, 10^{3}]$, optimizando la media geométrica en clasificación binaria estimada mediante una validación cruzada de 5 particiones repetida 2 veces (CV5x2). En el caso del SVM-RBF, el parámetro γ se seleccionó entre los valores [0,001, 0,005, 0,01, 0,05, 0,1, 1] usando el mismo procedimiento. Además de usar la media geométrica como medida a optimizar, en el caso de la regresión logística y del SVM-RBF se usaron costes para ponderar ambas clases de igual forma usando para ello el parámetro -w de LibLinear y LibSVM. Ambas decisiones tienen como objetivo obtener clasificadores buenos incluso en situaciones

en que las clases no estén balanceadas, situación habitual en problemas de cuantificación. Optimizando el acierto y sin usar costes, los resultados tanto del CC como del AC son muchos peores y la diferencia con respecto a los *ensembles* es aún mayor, debido a que el clasificador tiende a predecir en mayor medida la clase mayoritaria. EAC y EHDy usaron 30 modelos. Cada modelo se entrenó con una muestra del mismo tamaño que el conjunto original generada del siguiente modo: primero se escoge al azar una prevalencia entre el 5% y el 95% y, teniendo en cuenta esa prevalencia, se generan los ejemplos necesarios de cada clase (con remplazamiento para mantener constante $P(x|y)$). Se evita llegar cerca de los extremos 0% y 100% porque en esos casos el cálculo del tpr y el fpr puede ser problemático para EAC al contar con pocos ejemplos de una de las clases.

La Tabla 1 muestra los resultados de una CV5x2 de los cinco métodos usando RL como clasificador. Con cada partición de test se generaban 100 muestras con remplazamiento en las que la prevalencia de la clase positiva variaba entre el 0 % y el 100 %. Por tanto, cada número en la tabla representa la media de 1000 cuantificaciones. La medida de error de la tabla es el error cuadrático medio. Similares resultados se obtienen si analizásemos el error absoluto medio.

Como se puede apreciar, la aproximación trivial, clasificar y contar, es superada por el resto de métodos. Solamente en un caso (balance.2) CC es mejor que el resto de cuantificadores. Si atendemos a la comparativa entre AC/HDy y EAC/EHDy, estos últimos mejoran claramente a los primeros. Así en la comparación EAC vs AC, el primero gana en 26 ocasiones, pierde en 5 y hay un empate. Por su parte, EHDy supera en 22 ocasiones a HDy, pierde en 6 y hay 4 empates. Además las derrotas de las versiones con *ensembles* suelen producirse por márgenes muy pequeños en problemas con errores cuadráticos bajos, mientras en problemas más difíciles, suelen resultar vencedores por márgenes relativamente amplios en muchas ocasiones. Estos resultados son notables ya que en teoría AC debería, teóricamente, producir cuantificaciones perfectas, independientemente de la calidad del clasificador. La realidad es que en la práctica la corrección (2) suele funcionar bien en el entorno de la prevalencia del conjunto de entrenamiento original, pero peor para otros valores. Este detalle es uno de los justifica el uso de *ensembles*, ya que se obtienen modelos entrenados con diferentes prevalencias, que proporcionan estimaciones más precisas.

Estos resultados se pueden analizar estadísticamente de diferentes formas. En primer lugar, siguiendo [8], se realizó una comparación estadística en dos pasos: un test de Friedman rechaza la hipótesis nula de que todos los métodos obtienen un rendimiento igual y en segundo lugar, realizando comparaciones por pares mediante el test de Bergmann-Hommel con $\alpha = 0{,}05$, se observa que tanto EHDy como EAC son significativamente mejores que CC y AC; la diferencia entre EHDy y HDy no es significativa. Sin embargo, estas comparaciones se ven afectadas en gran medida por la inclusión de CC, ya que se está añadiendo un método que sistemáticamente pierde frente al resto, y sobre todo porque hay dependencias y correlaciones entre los métodos. En realidad, teniendo en cuenta que el objetivo del experimento era comparar los métodos basados en *ensembles* con sus homólogos, un test más apropiado es por ejemplo un test de Wilcoxon

Tabla 2. Resumen de resultados con distintos clasificadores base y distintas medidas. Los símbolos § y † indican diferencias significativa entre un *ensemble* y su homólgo usando un test de Bergmann-Hommel o un test de Wilcoxon, respectivamente

Clasificador	Medida	CC	AC	EAC	HDy	EHDy
Naïve Bayes	Ranking MSE	4.5625	3.7812 §†	1.9531	2.8438 §	**1.8594**
	Ranking MAE	4.3594	3.9219 §†	2.0469	2.7344 §	**1.9375**
Reg. Logística	Ranking MSE	4.5625	3.6875 §†	2.5938	2.4688 §	**1.6875**
	Ranking MAE	4.4219	3.7969 §†	2.6250	2.4688 §	**1.6875**
SVM-RBF	Ranking MSE	4.2031	2.7969	**2.0469**	3.2812	2.6719
	Ranking MAE	3.9375	2.9062	**2.1562**	3.2188	2.7812

de rangos con signo. Ese test muestra que EAC es significativamente mejor que AC ($p = 0{,}0001$) y lo mismo ocurre con EHDy respecto a HDy ($p = 0{,}0013$).

No se incluyen los resultados detallados usando NB y SVM-RBF por falta de espacio, pero la Tabla 2 muestra los resultados tanto para error cuadrático medio (MSE), como error absoluto medio (MAE). En el caso de NB, EAC y EHDy superan aún con más claridad a sus homólogos. Por ejemplo, en un test de Wilcoxon de rangos con signo EAC es significativamente mejor que AC ($p = 2{,}5e-5$) y lo mismo pasa con EHDy respecto a HDy ($p = 0{,}00023$). No ocurre lo mismo en el caso de SVM-RBF. Los *ensembles* son mejores, pero las diferencias no son significativas, salvo ligeramente en el caso de EHDy vs. HDy ($p = 0{,}076$). Las explicaciones que encontramos para este comportamiento son: 1) se trata de modelos más complejos, con más riesgo de sobre-ajuste, y 2) SVM no es puramente un clasificador probabilístico, sino que las probabilidades se obtienen tras un post-proceso. Eso provoca que los resultados sean más inestables, con modelos que pueden ser muy malos y otros mucho mejores para el problema. De hecho, de los tres clasificadores, la versión de CC con SVM-RBF es la peor.

5. Conclusiones

En este trabajo hemos estudiado el comportamiento de los *ensembles* en un problema en el que se asume un cambio en la distribución entre los datos de entrenamiento y el test. La idea central del artículo es precisamente generar las muestras usadas para entrenar los distintos modelos sabiendo el cambio esperado en la distribución, de forma que la diversidad se basa en adaptarse a dicho cambio. En concreto lo hemos aplicado al problema de la cuantificación binaria, en el que se conoce que la prevalencia de la clases varía, cambia $P(y)$, pero se asume que se mantiene constante $P(x|y)$. Las diferentes muestras se generan bajo esta asunción de forma que el meta-modelo está mejor entrenado para tratar conjuntos de test con distintas prevalencias. Los resultados experimentales prueban que los *ensembles* son capaces de mejorar el rendimiento de los cuantificadores en los que se basan.

El interés de este trabajo no radica únicamente en el hecho de presentar los primeros cuantificadores basados en *ensembles*, sino que esta misma idea es

también aplicable a otros problemas que presenten cambios en la distribución y donde se conozcan las características de dichos cambios. Además, en esa clase de problemas las opciones para desarrollar nuevos tipos de *ensembles* son aún mayores que para problemas de clasificación, tanto en los aspectos relativos a la diversidad adaptada al cambio esperado en la distribución, como sobre todo en las estrategias de combinación de los modelos. Aquí hemos usado simplemente la media, pero se pueden diseñar otros estrategias más sofisticadas que tengan en cuenta otros factores, por ejemplo, la prevalencia con la que se entrenó cada modelo o lo parecida que es su distribución respecto a la distribución del conjunto de test. Ambos aspectos abren vías de investigación interesantes para realizar nuevos trabajos en el campo del aprendizaje de *ensembles*.

Referencias

1. Jose Barranquero, Jorge Díez, and Juan José del Coz. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition*, 48(2):591–604, 2015.
2. Jose Barranquero, Pablo González, Jorge Díez, and Juan José Del Coz. On the study of nearest neighbor algorithms for prevalence estimation in binary problems. *Pattern Recognition*, 46(2):472–482, 2013.
3. Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. Quantification via probability estimators. In *IEEE International Conference on Data Mining (ICDM'10)*, pages 737–742, 2010.
4. Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.
5. Tom Fawcett and Peter Flach. A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning*, 58(1):33–38, 2005.
6. George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.
7. Giorgio Fumera and Fabio Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):942–956, 2005.
8. Santiago García and Francisco Herrera. An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.
9. Víctor. González-Castro, Rocio Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the hellinger distance. *Information Sciences*, 2012.
10. Ludmila Kuncheva. Classifier ensembles for changing environments. In *Multiple classifier systems*, pages 1–15. 2004.
11. Letizia Milli, Anna Monreale, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi, and Fabrizio Sebastiani. Quantification trees. In *IEEE International Conference on Data Mining (ICDM'13)*, pages 528–536, 2013.
12. J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
13. Joaquin Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.
14. Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.